Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 308340, 14 pages
doi:10.1155/2009/308340

Research Article

Modelling Errors in Automatic Speech Recognition for Dysarthric Speakers

Santiago Omar Caballero Morales and Stephen J. Cox

Speech, Language, and Music Group, School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK

Correspondence should be addressed to Santiago Omar Caballero Morales, s.caballero-morales@uea.ac.uk

Received 3 November 2008; Revised 27 January 2009; Accepted 24 March 2009

Recommended by Juan I. Godino-Llorente

Dysarthria is a motor speech disorder characterized by weakness, paralysis, or poor coordination of the muscles responsible for speech. Although automatic speech recognition (ASR) systems have been developed for disordered speech, factors such as low intelligibility and limited phonemic repertoire decrease speech recognition accuracy, making conventional speaker adaptation algorithms perform poorly on dysarthric speakers. In this work, rather than adapting the acoustic models, we model the errors made by the speaker and attempt to correct them. For this task, two techniques have been developed: (1) a set of "metamodels" that incorporate a model of the speaker's phonetic confusion matrix into the ASR process; (2) a cascade of weighted finite-state transducers at the confusion matrix, word, and language levels. Both techniques attempt to correct the errors made at the phonetic level and make use of a language model to find the best estimate of the correct word sequence. Our experiments show that both techniques outperform standard adaptation techniques.

1. Introduction

"Dysarthria is a motor speech disorder that is often associated with irregular phonation and amplitude, incoordination of articulators, and restricted movement of articulators" [1]. This condition can be caused by a stroke, cerebral palsy, traumatic brain injury (TBI), or a degenerative neurological disease such as Parkinson's Disease, or Alzheimer's Disease. The affected muscles by this condition may include the lungs, larynx, oropharynx and nasopharynx, soft palate, and articulators (lips, tongue, teeth, and jaw), and the degree to which these muscle groups are compromised determines the particular pattern of speech impairment [1].

Based on the presentation of symptoms, dysarthria is classified as *flaccid, spastic, mixed spastic-flaccid, ataxic, hyperkinetic,* and *hypokinetic* [2–4]. In all types of dysarthria, *phonatory dysfunction* is a frequent impairment and is difficult to assess because it often occurs along with other impairments affecting articulation, resonance, and respiration [2–6]. Particularly, six impairment features are related to phonatory dysfunction, reducing the speaker's intelligibility and altering naturalness of his/her speech [4, 7, 8].

(i) Monopitch: in all types of dysarthria.
(ii) Pitch level: in spastic and mixed spastic-flaccid.
(iii) Harsh voice: in all types of dysarthria.
(iv) Breathy voice: in flaccid and hypokinetic.
(v) Strained-strangled: in spastic and hyperkinetic.
(vi) Audible inspiration: in flaccid.

These features make the task of developing assistive Automatic Speech Recognition (ASR) systems for people with dysarthria very challenging. As a consequence of phonatory dysfunction, dysarthric speech is typically characterized by strained phonation, imprecise placement of the articulators and incomplete consonants closure. Intelligibility is affected when there is reduction or deletion of word-initial consonants [9]. Because of these articulatory deficits, the pronunciation of dysarthric speakers often deviates from that of nondysarthric speakers in several aspects: rate of speech is lower; segments are pronounced differently; pronunciation is less consistent; for longer stretches of speech, pronunciation can be even more varying due to fatigue [10]. Speaking rate, which is important for ASR performance, is affected by slow

pronunciation that produces prolonged phonemes. This can make a 1-syllable word to be interpreted as a 2-syllable word (*day* → *dial*), and words with long voiceless stops can be interpreted as two words because of the long silent occlusion phase in the middle of the target word (*before* → *be for*) [11].

The design of ASR systems for dysarthric speakers is difficult because they require different types of ASR depending on their particular type and level of disability [1]. Additionally, phonatory dysfunction and related impairments cause dysarthric speech to be characterized by phonetic distortions, substitutions, and omissions [12, 13] that decrease the speaker's intelligibility [1] and thus ASR performance. However it is important to develop ASR systems for dysarthric speakers because of the advantages they offer when compared with interfaces such as switches or keyboards. These may be more physically demanding and tiring [14–17] and as dysarthria is usually accompanied by other physical handicaps, impossible for them to use. Even with the speech production difficulties exhibited by many of these speakers, speech communication requires less effort and is faster than conventional typing methods [18], despite the difficulty of achieving robust recognition performance.

Experiments with commercial ASR systems have shown levels of recognition accuracy up to 90% for some dysarthric speakers with high intelligibility after a certain number of tests, although speakers with lower intelligibility did not achieve comparable levels of recognition accuracy [11, 19–22]. Most of the speakers involved in these studies presented individual error patterns, and variability in recognition rates was observed between test sessions and when trying different ASR systems. Usually these commercial systems require some speech samples from the speaker to adapt to his/her voice and thus increase recognition performance. However the system, which is trained on a normal speech corpus, is not expected to work well on severely dysarthric speech as adaptation techniques are insufficient to deal with gross abnormalities [16]. Moreover, it has been reported that recognition performance on such systems rapidly deteriorates for vocabulary sizes greater than 30 words, even for speakers with mild to moderate dysarthria [23].

Thus, research has concentrated on techniques to achieve more robust ASR performance. In [22], a system based on Artificial Neural Networks (ANNs) produced better results when compared with a commercial system, and outperformed the recognition of human listeners. In [10], the performance of HMM-based speaker-dependent (SD) and speaker-independent (SI) systems on dysarthric speech was evaluated. SI systems are trained on nondysarthric speech (as commercial systems above) and SD systems are trained on a limited amount of speech of the dysarthric speaker. The performance of the SD system was better than the SI's and the word error rates (WERs) obtained showed that ASR of dysarthric speech is certainly possible for low-perplexity tasks (with a highly constrained bigram language model).

The Center of Spoken Language Understanding [1] improved vowel intelligibility by the manipulation of a small set of highly relevant speech features. Although they limited themselves to studying consonant-vowel-consonant (CVC) contexts from a special purpose database, they significantly improved the intelligibility of dysarthric vowels from 48% to 54%, as evaluated by a vowel identification task using 64 CVC stimuli judged by 24 listeners. The ENABL Project ("ENabler for Access to computer-Based vocational tasks with Language and speech" [24, 25] was developed to provide access by voice via speech recognition to an engineering design system, ICAD. The baseline recognition engine was trained on nondysarthric speech (speaker-independent), and it was adapted to dysarthric speech using MLLR (Maximum Likelihood Linear Regression, see Section 2) [26]. This reduced the action error rate of the ICAD from 24.1% to 8.3%. However these results varied from speaker to speaker, and for some speakers the improvement was substantially greater than for others.

The STARDUST Project (Speech Training And Recognition for Dysarthric Users of Speech Technology) [16, 27–29] has developed speech technology for people with severe dysarthria. Among the applications developed, an ECS (Environmental Control System) was designed for home control with a small vocabulary speaker-dependent recognizer (10 words commands). The methodology for building the recognizer was adapted to deal with scarcity of training data and the increased variability of the material which was available. This problem was addressed by closing the loop between recognizer-training and user-training. They started by recording a small amount of speech data from the speaker, then they trained a recognizer using that data, and later used it to drive a user-training application, which allowed the speaker to practice to improve consistency of articulation. The speech-controlled ECS was faster to use than switch-scanning systems. Other applications from STARDUST are the following.

(i) STRAPTk (Speech Training Application Toolkit) [29], a system that integrates tools for speech analysis, exercise tasks, design, and evaluation of recognizers.

(ii) VIVOCA (Voice Input Voice Output Communication Aid) [30], which is aimed to develop a portable speech-in/speech-out communication aid for people with disordered or unintelligible speech. Another tool, the "Speech Enhancer" from Voicewave Technology Inc. [31], improves speech communication in real time for people with unclear speech and inaudible voice [32]. While VIVOCA recognizes disordered speech and resynthesises it in a normal voice, the Speech Enhancer does not recognize or correct speech distortions due to dysarthria.

A project at the University of Illinois is aimed to provide (1) a freely distributable multimicrophone, multicamera audiovisual database of dysarthric speech [33], and (2) programs and training scripts that could form the foundation for an open-source speech recognition tool designed to be useful for dysarthric speakers. In the University of Delaware, research has been done by the Speech Research Lab [34] to develop natural sounding software for speech synthesis (ModelTalker) [35], tools for articulation training for children (STAR), and a database of dysarthric speech [36].

As already mentioned, commercial "dictation" ASR systems have shown good performance for people with mild to moderate dysarthria [20, 21, 37], although these systems fail for speakers with more severe conditions [11, 22]. Variability in recognition accuracy, the speaker's inability to access the system by him/herself, restricted vocabulary, and continuous assistance and editing of words were evident in these studies. Although isolated-words recognizers have performed better than continuous speech recognizers, these are limited by their small vocabulary (10–78 possible words or commands), making them only suitable for "control" applications. For communication purposes, a continuous speech recognizer can be more suitable, and studies have shown that under some conditions a continuous system can perform better than a discrete system [37].

The motivation of our research is to develop techniques that could lead to the development of large-vocabulary ASR systems for speakers with different types of dysarthria, particularly when speech data for adaptation or training is small. In this paper, we describe two techniques that incorporate a model of the speaker's pattern of errors into the ASR process in such a way as to increase word recognition accuracy. Although these techniques have general application to ASR, we believe that they are particularly suitable for use in ASR of dysarthric speakers who have low intelligibility due, in some degree, to a limited phonemic repertoire [13], and the results presented here confirm this.

We continue in Section 1.1 by showing the pattern of errors caused on ASR due to the effects of a limited phonemic repertoire and thus expand on the effect of phonatory dysfunction on dysarthric speech. The description of our research starts in Section 2 with the details of the baseline system used for our experiments, the adaptation technique used for comparison, the database of dysarthric speech, and some initial word recognition experiments. In Section 3 the approach of incorporating information from the speaker's pattern of errors into the recognition process is explained. In Section 4 we present the first technique ("metamodels"), and in Section 5, results on word recognition accuracy when it is applied on dysarthric speech. Section 6 comments on the technique and motivates the introduction of a second technique in Section 7, which is based on a network of Finite-State Transducers (WFSTs). The results of this technique are presented in Section 8. Finally, conclusions and future work are presented in Section 9.

1.1. Limited Phonemic Repertoire. Among the identified factors that give rise to ASR errors in dysarthric speech [13], the most important are decreased intelligibility (because of substitutions, deletions, and insertions of phonemes) and limited phonemic repertoire, the latter leading to phoneme substitutions. To illustrate the effect of reduced phonemic repertoire, Figure 1 shows an example phoneme confusion matrix for a dysarthric speaker from the NEMOURS Database of Dysarthric Speech (described in Section 2). This confusion matrix is estimated by a speaker-independent ASR system, and so it may show confusions that would not actually be made by humans, and also spurious confusions

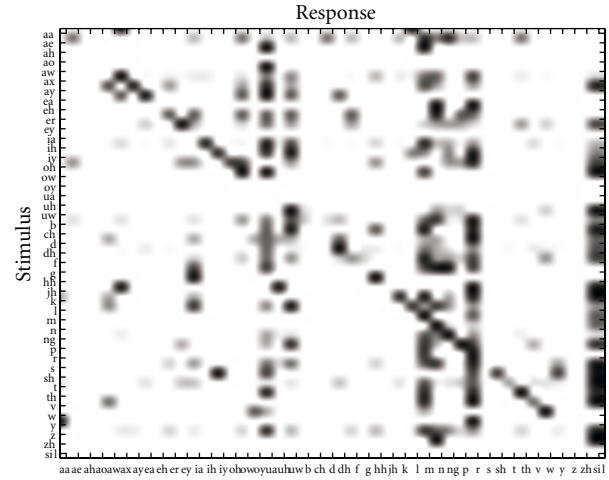

Figure 1: Phoneme confusion matrix from a dysarthric speaker.

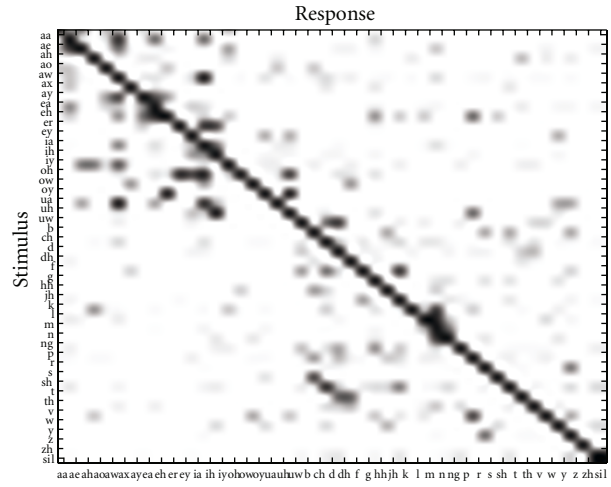

Figure 2: Phoneme confusion matrix from a normal speaker.

that are actually caused by poor transcription/output alignment (see Section 4.1). However, since we concerned with machine rather than human recognition here, we can make the following observations.

(i) A small set of phonemes (in this case the phonemes /ua/, /uw/, /m/, /n/, /ng/, /r/, and /sil/) dominates the speaker's output speech.

(ii) Some vowel sounds and the consonants /g/, /zh/, and /y/, are never recognized correctly. This suggests that there are some phonemes that the speaker apparently cannot enunciate at all, and for which he or she substitutes a different phoneme, often one of the dominant phonemes mentioned above.

These observations differ from the pattern of confusions seen in a normal speaker from the Wall Street Journal (WSJ) database [38], as shown in Figure 2. This confusion matrix shows a clearer pattern of correct recognitions and few confusions of vowels with consonants.

Most speaker adaptation algorithms are based on the principle that it is possible to apply a set of transformations to the parameters of a set of acoustic models of an "average" voice to move them closer to the voice of an individual. Whilst this has been shown to be successful for normal speakers, it may be less successful in cases where the phoneme uttered is not the one that was intended but is substituted by a different phoneme or phonemes, as often happens in dysarthric speech. In this situation, we argue that a more effective approach is to combine a model of the substitutions likely to have been made by the speaker with a language model to infer what was said. So rather than attempting to adapt the system, we model the insertion, deletion, and substitution errors made by a speaker and attempt to correct them.

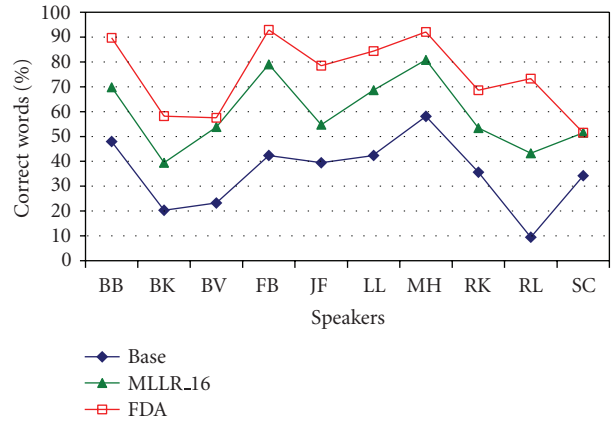

Figure 3: Comparison of recognition performance: human assessment (FDA), unadapted (BASE) and adapted (MLLR_16) SI models.

2. Speech Data, Baseline Recognizer, and Adaptation Technique

Our speaker-independent (SI) speech recognizer was built with the HTK Toolkit [39] using the data from 92 speakers in set *si_tr* of the Wall Street Journal (WSJ) database [38]. A Hamming window of 25 milliseconds moving at a frame rate of 10 milliseconds was applied to the waveform data to convert it to 12 MFCCs (using 26 filterbanks), and energy, delta, and acceleration coefficients were added. The resulting data was used to construct 45 monophone acoustic models. The monophone models had a standard three state left-to-right topology with eight mixture components per state. They were trained using standard maximum-likelihood techniques, using the routines provided in HTK.

The dysarthric speech data was provided by the NEMOURS Database [36]. This database is a collection of 814 short sentences spoken by 11 speakers (74 sentences per speaker) with varying degrees of dysarthria (data from only 10 speakers was used as some data is missing for one speaker). The sentences are nonsense phrases that have a simple syntax of the form "the X is Y the Z", where X and Z are usually nouns and Y is a verb in present participle form (for instance, the phrases "The shin is going the who", "The inn is heaping the shin", etc.). Note that although each of the 740 sentences is different, the vocabulary of 112 words is shared.

Speech recognition experiments were implemented by using the baseline recognizer on the dysarthric speech. For these experiments, a word-bigram language model was estimated from the (pooled) 74 sentences provided by each speaker.

The technique used for the speaker adaptation experiments was MLLR (Maximum Likelihood Linear Regression) [26]. A two-pass MLLR adaptation was implemented as described in [39], where a global adaptation is done first by using only one class. This produces a global-input transformation that can be used to define more specific transforms to better adapt the baseline system to the speaker's voice. Dynamic adaptation is then implemented by using a regression class tree with 32 terminal nodes or base classes.

From the complete set of 74 sentences per speaker, 34 sentences were used for adaptation and the remaining 40 for testing. The set of 34 was divided into sets to measure the performance of the adapted baseline system when using a different amount of adaptation data. Thus adaptation was implemented using 4, 10, 16, 22, 28, and 34 sentences. For future reference, the baseline system adapted with X sentences will be termed as MLLR_X and the baseline without any adaptation as BASE.

Table 1 shows the number of MLLR transform classes (XFORMS) for the 10 dysarthric speakers used in these experiments using different amounts of adaptation data. For comparison purposes, Table 2 shows the same for ten speakers selected randomly from the *si_dt* set of the WSJ database using similar sets of adaptation data. In both cases, the number of transforms increases as more data is available. The mean number of transforms (Mean_XFORMS) is similar for both sets of speakers, but the standard deviation (STDEV) is higher for dysarthric speakers. This shows that within dysarthric speakers there are more differences and variability than within normal speakers, which may be caused by individual patterns of phonatory dysfunction.

An experiment was done to compare the performance of the baseline and MLLR-adapted recognizer (using 16 utterances for adaptation) with a human assessment of the dysarthric speakers used in this study. Recognition was performed with a grammar scale factor and word insertion penalty as described in [39].

Figure 3 shows the intelligibility of each of the dysarthric speakers as measured using the Frenchay Dysarthria Assessment (FDA) test in [36], and the recognition performance (% word correct) when tested on the unadapted baseline system (BASE) and the adapted models (MLLR_16). The correlation between the FDA performance and the recognizer performance is 0.67 (unadapted models) and 0.82 (adapted). Both are significant at the 1% level, which gives some confidence that the recognizer displays a similar performance trend when exposed to different degrees of dysarthric speech as humans.

Table 1: MLLR transforms for dysarthric speakers.

Adaptation data	Dysarthric speakers										Mean_XFORMS	STDEV
	BB	BK	BV	FB	JF	LL	MH	RK	RL	SC		
4	0	4	1	2	2	1	1	1	5	3	2	1.6
10	3	10	5	4	5	4	4	3	8	7	5	2.3
16	5	11	6	7	7	5	5	5	11	9	7	2.4
22	7	11	7	9	10	9	8	6	11	11	9	1.9
28	9	11	9	9	10	10	10	8	11	12	10	1.2
34	10	11	10	9	11	11	10	9	11	12	10	1.0

Table 2: MLLR transforms for normal speakers.

Adaptation data	Normal (WSJ) speakers										Mean_XFORMS	STDEV
	C31	C34	C35	C38	C3C	C40	C41	C42	C45	C49		
5	5	4	6	5	3	3	5	5	4	3	4	1.1
10	8	7	8	6	7	8	6	6	7	6	7	0.9
15	11	10	9	9	9	11	9	8	10	9	10	1.0
20	12	12	12	11	10	12	11	9	12	11	11	1.0
30	13	13	13	12	11	13	13	11	13	12	12	0.8

3. Incorporating a Model of the Confusion Matrix into the Recognizer

We suppose that a dysarthric speaker wishes to utter a word-sequence W that can be transcribed as a phone sequence P. In practice, he or she utters a different phone sequence $\widetilde{P}$. Hence the probability of the acoustic observations O produced by the speaker given W can be written as

$$\Pr(O \mid W) = \Pr(O \mid P) = \sum_{\widetilde{P}} \Pr\left(O \mid P, \widetilde{P}\right) \Pr\left(\widetilde{P} \mid P\right). \quad (1)$$

However, once $\widetilde{P}$ is known, there is no dependence of O on P, so we can write

$$\Pr(O \mid W) = \sum_{\widetilde{P}} \Pr\left(O \mid \widetilde{P}\right) \Pr\left(\widetilde{P} \mid P\right). \quad (2)$$

Hence the probability of a particular word sequence W^* with associated phone sequence P^* is

$$\Pr(P^* \mid O) = \frac{\Pr\left(O \mid \widetilde{P}\right) \Pr(P^*)}{\Pr(O)} \quad (3)$$

$$= \frac{\Pr(P^*) \sum_{\widetilde{P}} \Pr\left(O \mid \widetilde{P}\right) \Pr\left(\widetilde{P} \mid P\right)}{\Pr(O)}. \quad (4)$$

In the usual way, we can drop the denominator of (4), as it is common to all W sequences. Furthermore, we can approximate

$$\sum_{\widetilde{P}} \Pr\left(O \mid \widetilde{P}\right) \Pr\left(\widetilde{P} \mid P\right) \approx \max_{\widetilde{P}} \Pr\left(O \mid \widetilde{P}\right) \Pr\left(\widetilde{P} \mid P\right) \quad (5)$$

which will be approximately correct when a single phone sequence dominates. The observed phone sequence from the dysarthric speaker, $\widetilde{P}^*$, is obtained as

$$\widetilde{P}^* = \operatorname*{argmax}_{\widetilde{P}} \Pr\left(O \mid \widetilde{P}\right) \quad (6)$$

from a phone recognizer, which also provides the term $\Pr(O \mid \widetilde{P}^*)$. Hence the most likely phone sequence is given as

$$P^* = \operatorname*{argmax}_{\widetilde{P}} \Pr(P) \Pr\left(O \mid \widetilde{P}^*\right) \Pr\left(\widetilde{P}^* \mid P\right), \quad (7)$$

where it is understood that P^* ranges over all valid phone sequences defined by the dictionary and the language model. If we now make the assumption of conditional independence of the individual phones in the sequences P^* and $\widetilde{P}^*$, we can write

$$W^* = \operatorname*{argmax}_{P} \prod_j \Pr\left(p_j\right) \Pr\left(\widetilde{p}_j^* \mid p_j\right), \quad (8)$$

where p_j is the jth phoneme in the postulated phone sequence P, and $\widetilde{p}_j^*$ the jth phoneme in the decoded sequence $\widetilde{P}^*$ from the dysarthric speaker. Equation (8) indicates that the most likely word sequence is the sequence that is most likely given the observed phone sequence from the dysarthric speaker. The term $\Pr(\widetilde{p}_j^* \mid p_j)$ is obtained from a confusion matrix for the speaker.

The overall procedure to use the estimates of $\Pr(\widetilde{p}_j^* \mid p_j)$ into the recognition process is presented in Figure 4. A set of training sentences (as described in Section 2) is used to estimate $\Pr(\widetilde{p}_j^* \mid p_j)$ and identify patterns of deletions/insertions of phonemes. This information is modelled by our two techniques that will be presented in Sections 4 and 7. Evaluation is performed when $\widetilde{P}^*$ (which now is obtained from test speech) is decoded by using the "trained" techniques into sequences of words W^*. The correction process is done at the phonemic level, and by incorporating a word language model a more accurate estimate of W is obtained.

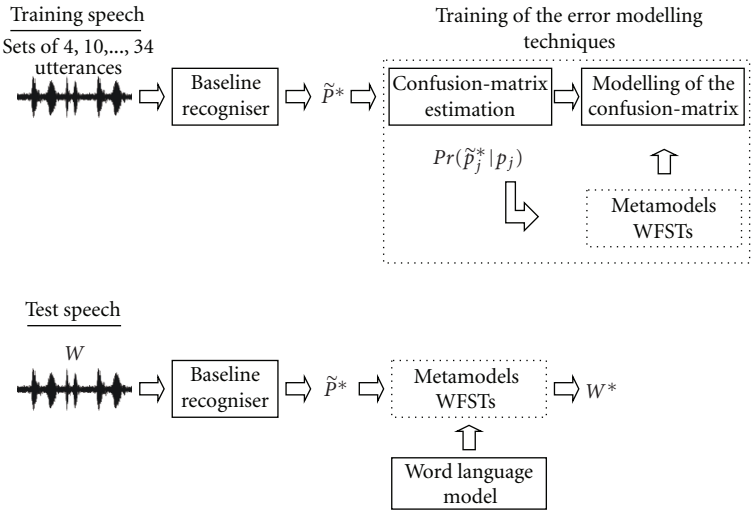

Figure 4: Diagram of the correction process.

Table 3: *Upper pair*: alignment of transcription and recognized output using *HResults*; *Lower pair*: same, using the improved aligner.

P:				dh	ax	sh	uw	ih	z	b	ea	r	ih	ng	dh	ax	b	ey	dh
$\widetilde{P}^*$:	dh	ax	ng	dh	ax	y	ua	ng	dh	ax	b	l	ih	ng	dh	ax	b		uw
P:	dh	ax		sh	uw		ih	z	b	ea		r	ih	ng	dh	ax	b	ey	dh
$\widetilde{P}^*$:	dh	ax	ng	dh	ax	y	ua	ng	dh	ax	b	l	ih	ng	dh	ax	b	uw	

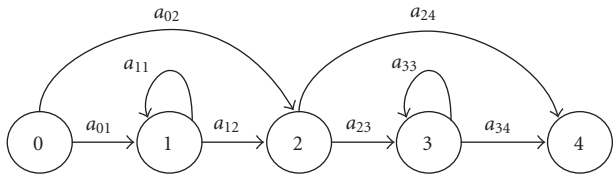

Figure 5: Metamodel of a phoneme.

4. First Technique: Metamodels

In practice, it is too restrictive to use only the confusion matrix to model $\Pr(\widetilde{p}_j^* \mid p_j)$ as this cannot model insertions well. Instead, a hidden Markov model (HMM) is constructed for each of the phonemes in the phoneme inventory. We term these HMMs *metamodels* [40]. The function of a metamodel is best understood by comparison with a "standard" acoustic HMM: a standard acoustic HMM estimates $\Pr(O' \mid p_j)$, where O' is a subsequence of the complete sequence of observed acoustic vectors in the utterance, O, and p_j is a postulated phoneme in P. A metamodel estimates $\Pr(\widetilde{P}' \mid p_j)$, where $\widetilde{P}'$ is a subsequence of the complete sequence of observed (decoded) phonemes in the utterance $\widetilde{P}$.

The architecture of the metamodel of a phoneme is shown in Figure 5. Each state of a metamodel has a discrete probability distribution over the symbols for the set of phonemes, plus an additional symbol labelled DELETION. The central state (2) of a metamodel for a certain phoneme models correct decodings, substitutions, and deletions of this phoneme made by the phone recognizer. States 1 and 3 model (possibly multiple) insertions before and after the phoneme. If the metamodel were used as a generator, the output phone sequence produced could consist of, for example,

(i) a single phone which has the same label as the metamodel (a correct decoding) or a different label (a substitution);

(ii) a single phone labelled DELETION (a deletion);

(iii) two or more phones (one or more insertions).

As an example of the operation of a metamodel, consider a hypothetical phoneme that is always decoded correctly without substitutions, deletions, or insertions. In this case, the discrete distribution associated with the central state would consist of zeros except for the probability associated with the symbol for the phoneme itself, which would be 1.0. In addition, the transition probabilities a_{02} and a_{24} would be set to 1.0 so that no insertions could be made. When used as a generator, this model can produce only one possible phoneme sequence: a single phoneme which has the same label as the metamodel.

We use the reference transcription P of a training set utterance to enable us to concatenate the appropriate sequence of phoneme metamodels for this utterance. The associated recognition output sequence $\widetilde{P}^*$ for the utterance is obtained from the phoneme transcription of the word sequences decoded by a speech recognizer and is used to

train the parameters of the metamodels in this utterance. Note that the speech recognizer itself can be built using unadapted or MLLR adapted phoneme models. By using embedded reestimation over the $\{P, \widetilde{P}^*\}$ pairs of all the utterances, we can train the complete set of metamodels. In practice, the parameters formed, especially the probability distributions, are sensitive to the initial values to which they are set, and it is essential to "seed" the probabilities of the distributions using data obtained from an accurate alignment of P and $\widetilde{P}^*$ for each training-set sentence. After the initial seeding is complete, the parameters of the metamodels are reestimated using embedded reestimation as described above. Before recognition, the language model is used to compile a "metarecognizer" network, which is identical to the network used in a standard word recognizer except that the nodes of the network are the appropriate metamodels rather than the acoustic models used by the word recognizer. At recognition time, the output phoneme sequence $\widetilde{P}^*$ is passed to the metarecognizer to produce a set of word hypotheses.

4.1. Improving Alignment for Confusion Matrix Estimation. Use of a standard dynamic programming (DP) tool to align two symbol strings (such as the one available in the *HResults* routine in the HTK package [39]) can lead to unsatisfactory results when a precise alignment is required between P and $\widetilde{P}^*$ to estimate a confusion matrix, as is the case here. This is because these alignment tools typically use a distance measure which is "0" if a pair of symbols are the same, "1" otherwise. In the case of *HResults*, a correct match has a score of "0", an insertion and a deletion carry a score of "7", and a substitution a score of "10" [39]. To illustrate this, consider the top alignment in Table 3, which was made using *HResults*. It is not a plausible alignment, because

(i) the first three phones in the recognized output are unaligned and so must be regarded as insertions;

(ii) the fricative /sh/ in the transcription has been aligned to the vocalic /y/;

(iii) the sequence /b ea/ in the transcription has been aligned to the sequence /ax b/.

In the lower alignment in Table 3, these problems have been rectified, and a more plausible alignment results. This alignment was made using a DP matching algorithm in which the distance $D(\widetilde{p}_j^*, p_j)$ between a phone in the reference transcription P and a phone in the recognition output $\widetilde{P}^*$ considers a similitude score given by the empirically derived expression:

$$\mathrm{Sim}\left(\widetilde{p}_j^*, p_j\right) = 5\,\mathrm{Pr}_{\mathrm{SI}}\left(\widetilde{q}_j^* \mid q_j\right) - 2, \tag{9}$$

where $\mathrm{Pr}_{\mathrm{SI}}(\widetilde{q}_j^* \mid q_j)$ is a speaker-independent confusion matrix pooled over 92 WSJ speakers and is estimated by a DP algorithm that uses a simple aligner (e.g., HResults). Hence, a pair of phonemes that were always confused is assigned a score of +3, and a pair that is never confused is assigned a score of −2. The effect of this is that the DP algorithm prefers to align phoneme pairs that are more likely to be confused.

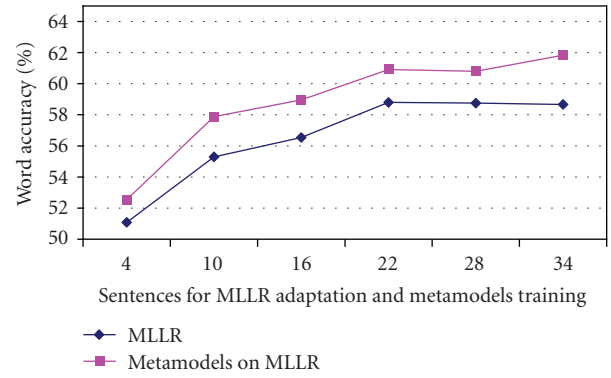

Figure 6: Mean word recognition accuracy of the adapted models and the metamodels across all dysarthric speakers.

5. Results of the Metamodels on Dysarthric Speakers

Figure 6 shows the results of the metamodels on the phoneme strings from the MLLR adapted acoustic models. When a very small set of sentences, for example, four, is used for training of the metamodels, it is possible to get an improvement of approximately 1.5% over the MLLR adapted models. This gain in accuracy increases as the training/adaptation data is increased, obtaining an improvement of almost 3% when all 34 sentences are used. The matched pairs test described in [41] was used to test for significant differences between the recognition accuracy using metamodels and the accuracy obtained with MLLR adaptation when a certain number of sentences were available for metamodel training. The results with the associated P-values are presented in Table 4. In all the cases, metamodels improve MLLR adaptation with P-values less than .01 and .05. Note that the metamodels trained with only four sentences (META_04) decrease the number of word errors from 1174 (MLLR_04) to 1139.

5.1. Low and High Intelligibility-Speakers. Low intelligibility-speakers were classified as those with low recognition performances using the unadapted and adapted models. As shown in Figure 3, automatic recognition followed a similar trend to human recognition (as scored by the FDA intelligibility test). So in the absence of a human assessment test, it is reasonable to classify a speaker's intelligibility based on their automatic recognition performance.

The set of speakers was divided into two equal-sized groups: high intelligibility (BB, FB, JF, LL, and MH), and low intelligibility (BK, BV, RK, RL, and SC). In Figure 7 the results for all low intelligibility speakers are presented. There is an overall improvement of about 5% when using different training sets. However for speakers with high intelligibility, there is no improvement over MLLR, as shown in Figure 8.

Table 4: Comparison of statistical significance of results over all dysarthric speakers.

System	Errors	P
MLLR_04	1174	.00168988
META_04	1139	
MLLR_10	1073	.0002459
META_10	1036	
MLLR_16	1043	.00204858
META_16	999	
MLLR_22	989	.0000351
META_22	941	
MLLR_28	990	.00240678
META_28	952	
MLLR_34	992	.00000014
META_34	924	

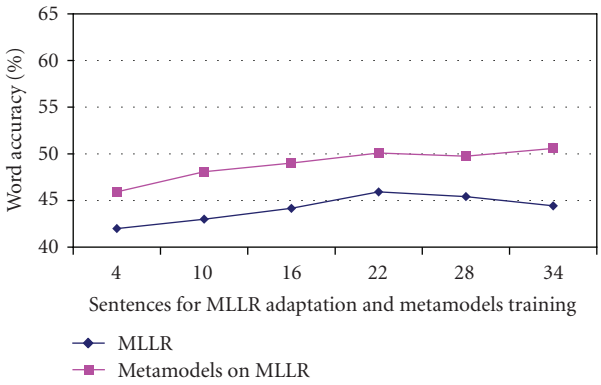

Figure 7: Mean word recognition accuracy of the adapted models and the metamodels across all low intelligibility dysarthric speakers.

These results indicate that the use of metamodels is a significantly better approach to ASR than speaker adaptation in cases where the intelligibility of the speaker is low and only a few adaptation utterances are available, which are two important conditions when dealing with dysarthric speech. We believe that the success of metamodels in increasing performance for low-intelligibility speakers can be attributed to the fact that these speakers often display a confusion matrix that is similar to the matrix shown in Figure 1, in which a few phonemes dominate the speaker's repertoire. The metamodels learn the patterns of substitution more quickly than the speaker adaptation technique, and hence perform better even when only a few sentences are available to estimate the confusion matrix.

6. Limitations of the Metamodels

As presented in Section 5, we had some success using the metamodels on dysarthric speakers. However the experiments showed that they suffered from two disadvantages.

(1) The models had a problem dealing with deletions. If the metamodel network defining a legal sequence of words is defined in such a way that it is possible to traverse it by "skipping" every metamodel, the decoding algorithm fails because it is possible to traverse the complete network of HMMs without absorbing a single input symbol. We attempted to remedy this problem by adding an extra "deletion" symbol (see Section 4), but as this symbol could potentially substitute every single phoneme in the network, it led to an explosion in the size of the dictionary, which was unsatisfactory.

(2) The metamodels were unable to model specific phone sequences that were output in response to individual phone inputs. They were capable of outputting sequences, but the symbols (phones) in these sequences were conditionally independent, and so specific sequences cannot be modelled.

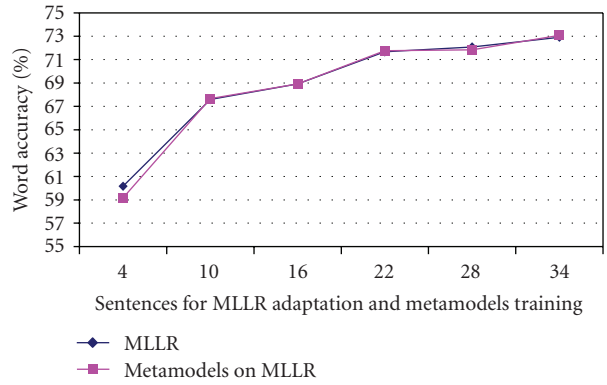

Figure 8: Mean word recognition accuracy of the adapted models and the metamodels across all high intelligibility dysarthric speakers.

A network of Weighted Finite-State Transducers (WFSTs) [42] is an attractive alternative to metamodels for the task of estimating W from $\tilde{P}^*$. WFSTs can be regarded as a network of automata. Each automaton accepts an input symbol and outputs one of a finite set of outputs, each of which has an associated probability. The outputs are drawn (in this case) from the same alphabet as the input symbols and can be single symbols, sequences of symbols, or the deletion symbol ε. The automata are linked by a set (typically sparse) of arcs and there is a probability associated with each arc.

These transducers can model the speaker's phonetic confusions. In addition, a cascade of such transducers can model the mapping from phonemes to words, and the mapping from words to a word sequence described by a grammar.

The usage proposed here complements and extends the work presented in [43], in which WFSTs were used to correct phone recognition errors. Here, we extend the technique to convert noisy phone strings into word sequences.

7. Second Technique: Network of Weighted Finite-State Transducers

As shown in, for instance, [42, 44], the speech recognition process can be realised as a cascade of WFSTs. In this

approach, we define the following transducers to decode $\widetilde{P}^*$ into a sequence of words W^*.

(1) C, the confusion matrix transducer, which models the probabilities of phoneme insertions, deletions, and substitutions.

(2) D, the dictionary transducer, which maps sequences of decoded phonemes from $\widetilde{P}^* \circ C$ into words in the dictionary.

(3) G, the language model transducer, which defines valid sequences of words from D.

Thus, the process of estimating the most probable sequence of words W^* given $\widetilde{P}^*$ can be expressed as

$$W^* = \tau^*\left(\widetilde{P}^* \circ C \circ D \circ G\right), \tag{10}$$

where τ^* denotes the operation of finding the most likely path through a transducer and $\circ$ denotes composition of transducers [42]. Details of each transducer used will be presented in the following sections.

7.1. Confusion Matrix Transducer (C). In this section, we describe the formation of the confusion matrix transducer C. In Section 3, we defined $\widetilde{p}_j^*$ as the jth phoneme in $\widetilde{P}^*$ and p_j as the jth phoneme in P, where $\Pr(\widetilde{p}_j^* \mid p_j)$ is estimated from the speaker's confusion matrix, which is obtained from an alignment of many sequences of $\widetilde{P}^*$ and P. While single substitutions are modelled in the same way by both, metamodels and WFSTs, insertions and deletions are modelled in a different way, thus taking advantage of the characteristics of the WFSTs. Here, the confusion matrix transducer C can map single and multiple phoneme insertions and deletions.

Consider Table 5, which shows an alignment from one of our experiments. The top row of phone symbols represents the transcription of the word sequence and the bottom row the output from the speech recognizer. It can be seen that the phoneme sequence /b aa/ is deleted after /ax/, and this can be represented in the transducer as a multiple substitution/insertion: /ax/ → /ax b aa/. Similarly the insertion of /ng dh/ after /ih/ is modeled as /ih ng dh/ → /ih/. The probabilities of these multiple substitutions/insertions/deletions are estimated by counting. In cases where a multiple insertion or deletion is made of the form A → /B C/, the appropriate fraction of the unigram probability mass $\Pr(\text{A} \rightarrow \text{B})$ is subtracted and given to the probability $\Pr(\text{A} \rightarrow /\text{B C}/)$, and the same process is used for higher-order insertions or deletions.

A fragment of the confusion matrix transducer that represents the alignment of Table 5 is presented in Figure 9. For computational convenience, the weight for each confusion in the transducer is represented as $-\log \Pr(\widetilde{p}_j^* \mid p_j)$. In practice, we have found it convenient to build an initial set of transducers directly from the speaker's "unigram" confusion matrix, which is estimated using each transcription/output alignment pair available from that speaker, and then to add extra transducers that represent

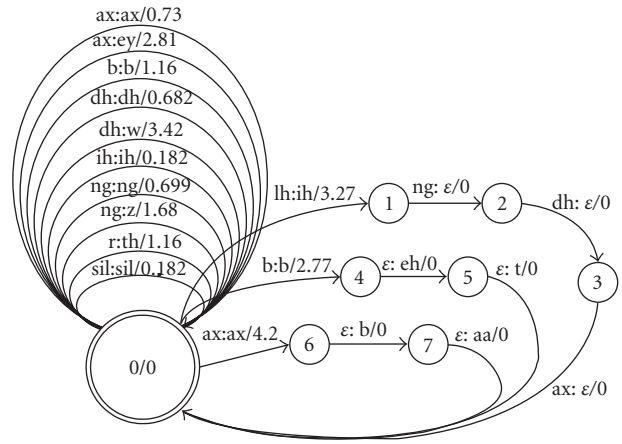

Figure 9: Example of the confusion matrix transducer C.

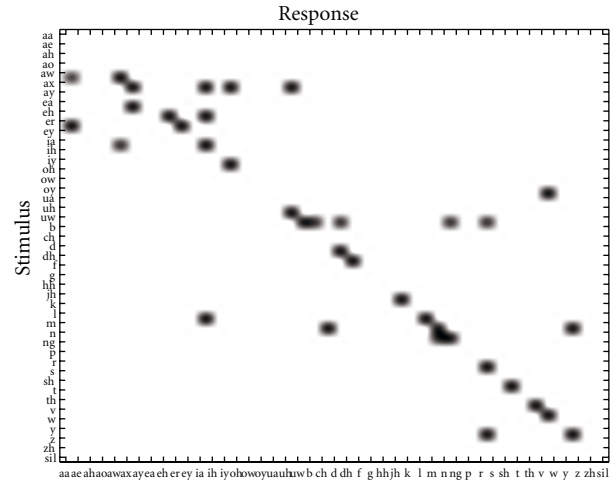

Figure 10: Sparse confusion matrix for C.

multiple substitution/insertion/deletions. The complete set of transducers are then determinized and minimized, as described in [42]. The result of these operations is a single transducer for the speaker.

One problem encountered when limited training data is available from speakers is that some phonemes are never decoded during the training phase, and therefore it is not possible to make any estimate of $\Pr(\widetilde{p}_j^* \mid p_j)$. This is shown in Figure 10, which shows a confusion matrix estimated from a single talker using only four sentences. Note that the columns are the response and the rows are the stimulus in this matrix, and so blank columns are phonemes that have never been decoded. We used two techniques to smooth the missing probabilities.

7.2. Base Smoothing. It is essential to have a nonzero value for every diagonal element of a confusion matrix to enable the decoding process to work using an arbitrary language model. One possibility is to set all diagonal elements for which no data exists to 1.0, that is, to assume that the associated phone is always correctly decoded. However, if the estimate of the

Table 5: Alignment of transcription P and recognized output $\widetilde{P}^*$.

P:	ax	b	aa	th	ih			ax	z	w	ey		ih	ng	dh	ax	b	eh	t
$\widetilde{P}^*$:	ax			r	ih	ng	dh	ax	ng	dh	ax	l	ih	ng	dh	ax	b		

overall probability of error of the recognizer on this speaker is p, a more robust estimate is to set any unseen diagonal elements to p, and we begin by doing this. We then need to decide how to assign nondiagonal probabilities for unseen confusions. We do this by "stealing" a small proportion of the probability mass on the diagonal and redistributing it along the associated row. This is equivalent to assigning a proportion of the probability of correctly decoded phonemes to as yet unseen confusions. The proportion of the diagonal probability that is used to estimate these unseen confusions depends on the amount of data from the speaker: clearly, as the data increases, the confusion probability estimates become more accurate and it is not appropriate to use a large proportion. Some experimentation on our data revealed that redistributing approximately 20% of the diagonal probability to unseen confusions worked well.

7.3. SI Smoothing. The "base" smoothing described in Section 7.2 could be regarded as "speaker-dependent" (SD) in that it uses the (sparse) confusion estimates made from the speaker's own data to smooth the unseen confusions. However, these estimates are likely to be noisy, so we add another layer of smoothing using the speaker-independent (SI) confusion matrix whose elements are well-estimated from 92 speakers of the WSJ database (see Section 2). The influence of this confusion matrix on the speaker-dependent matrix is controlled by a mixing factor *lambda*. Defining the elements of the SI confusion matrix as $\widetilde{q}_j^*$ and q_j (see Section 4.1) the resulting joint confusion matrix can be expressed as

$$\begin{aligned} C_{\text{joint}} &= \lambda \text{SI} + (1-\lambda)\text{SD} \\ &= \lambda \text{Pr}_{\text{SI}}\left(\widetilde{q}_j^* \mid q_j\right) + (1-\lambda)\text{Pr}\left(\widetilde{p}_j^* \mid p_j\right). \end{aligned} \tag{11}$$

The effect of both the base smoothing and the SI smoothing on the sparse confusion matrix of Figure 10 can be seen in Figure 11. The effect of λ on the mean word accuracy across all dysarthric speakers is shown in Figure 14.

7.4. Dictionary and Language Model Transducer ($\mathbf{D}, \mathbf{G}$). The transducer D maps sequences of phonemes into valid words. Although other work has investigated the possibility of using WFSTs to model pronunciation in this component [45], in our study, the pronunciation modelling is done by the transducer C. A small fragment of the dictionary entries is shown in Figure 12(a), where each sequence of phonemes that forms a word is listed as an FST. The minimized union of all these word entries is shown in Figure 12(b). The single and multiple pronunciations of each word were taken from the British English BEEP pronouncing dictionary [39]. The

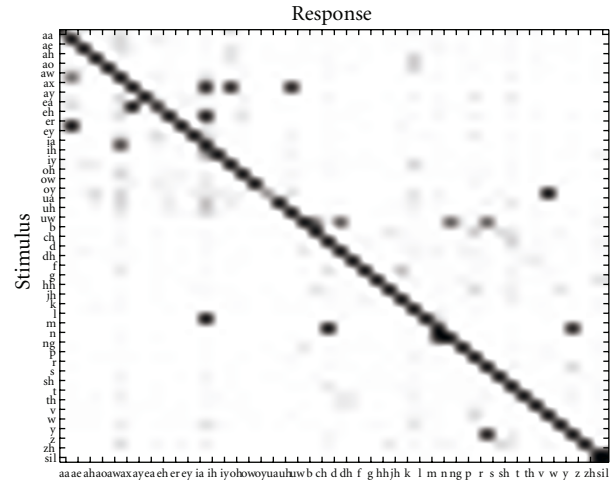

Figure 11: SI Smoothing of C, with $\lambda = 0.25$.

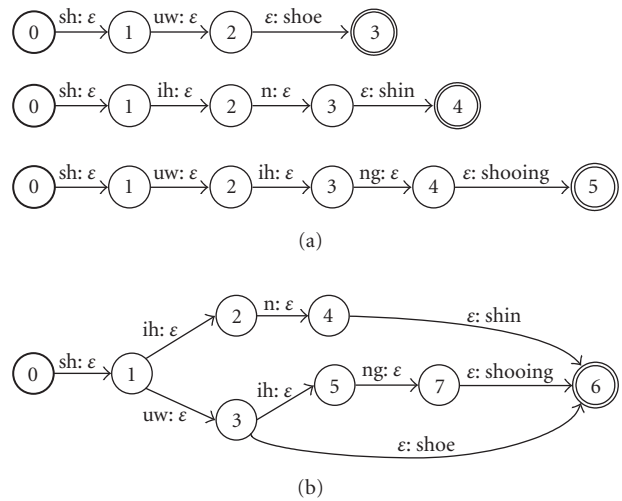

Figure 12: Example of the dictionary transducer D.

language model transducer consisted of a word bigram, as used in the metamodels, but now represented as a WFST. HLStats [39] was used to estimate these bigrams which were then converted into a format suitable for using in a WFST. A fragment of the word bigram FST G is shown in Figure 13. Note that the network of Figure 13 allows sequences of the form "the X is Y the Z" (see Section 2) to be recognized explicitly, but an arbitrary word bigram grammar can be represented using one of these transducers.

All three transducers used in these experiments were determinized and minimized in order to make execution more efficient.

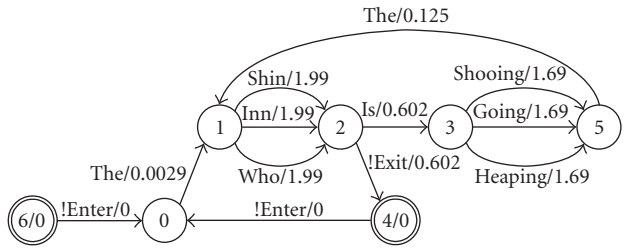

Figure 13: Example of the language model transducer G.

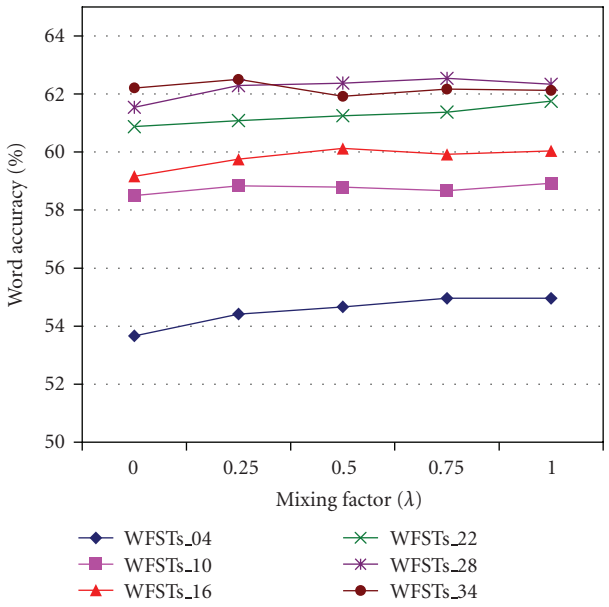

Figure 14: Mean across all dysarthric speakers: comparison of WFSTs performance for different values of λ.

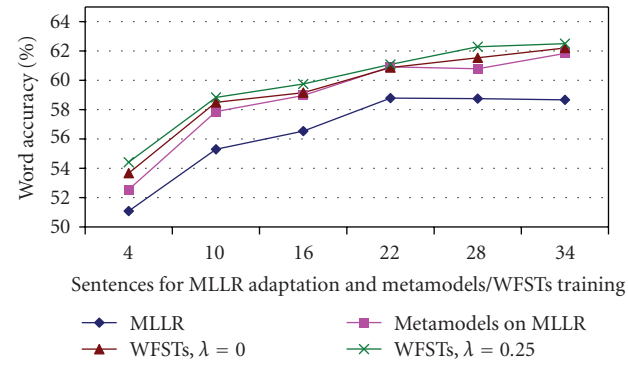

Figure 15: Mean across all dysarthric speakers: comparison of % word accuracy for different techniques.

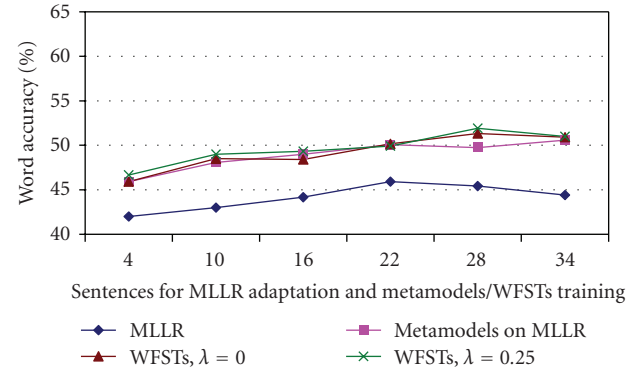

Figure 16: Mean word recognition accuracy of the adapted models, the metamodels, and the WFSTs across all low intelligibility dysarthric speakers.

8. Results of the WFST Approach on Dysarthric Speakers

The FSM Library [42, 46] from AT&T was used for the experiments with WFSTs. Figure 15 shows the mean word accuracies across all the dysarthric speakers for different amounts of adaptation data and using different decoding techniques. The Figure shows clearly the gain in performance given by the WFSTs over both MLLR and the metamodels, where the SI Smoothing increases the WFSTs performance over the Base Smoothing.

Note that Figure 15 shows results for only two values of λ: $\lambda = 0$ (Base Smoothing only) and SI Smoothing with $\lambda = 0.25$, since the variation in performance for values of λ above 0.25 is small, as observed in Figure 14.

When the WFSTs are trained with four and 22 utterances (WFSTs_04, WFSTs_22), best performance is obtained with $\lambda = 1$. WFSTs trained with 10 and 34 reach the maximum with $\lambda = 0.25$, while with 16 and 28 the maximum is obtained with $\lambda = 0.50$. However, the variation in performance is small for $\lambda > 0.25$ for most cases. It is important to mention that the mixing factor is applied to the unigram probability mass (see Section 3), which in turn affects the probability of insertion/deletion of any sequence of phonemes associated with that unigram. These sequences are still estimated from the data provided by the speaker, and thus are considered even when only the speaker-independent estimates are used ($\lambda = 1$).

8.1. Low and High Intelligibility Speakers. By separating the speakers into high and low intelligibility groups, as done in Section 5.1, a more detailed comparison of performance can be presented. In Figure 16, for low intelligibility speakers, the WFSTs with $\lambda = 0.25$ show a significant gain in performance over the metamodels when 4, 10, and 28 sentences are used for training. The gain in recognition accuracy is also evident for high intelligibility speakers, as shown in Figure 17. Figure 17 is encouraging because, as commented in Section 5.1, the modelling using metamodels did not achieve improvements on high intelligibility speakers. Hence WFSTs may be a useful technique to improve performance of recognition of normal speech.

9. Summary and Conclusions

We have argued that in the case of dysarthric speakers, who have a limited phonemic repertoire and thus consistently

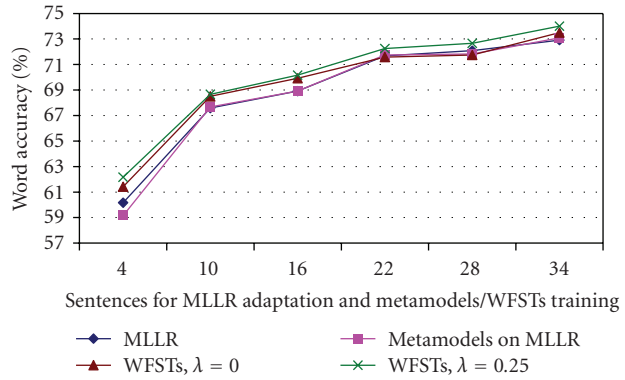

Figure 17: Mean word recognition accuracy of the adapted models, the metamodels, and the WFSTs across all high intelligibility dysarthric speakers.

substitute certain phonemes for others, modelling and correcting the errors made by the speaker under the guidance of a language model is a more effective approach than adapting acoustic models in the way that is effective for normal speakers. Our first system proposed the use of a technique called metamodels, which are HMM-like stochastic models that incorporate a model of a speaker's confusion matrix into the decoding process. Results obtained using metamodels showed a statistically significant improvement over the standard MLLR algorithm when the speech has low intelligibility and there is limited adaptation data available for a speaker, two conditions that are often met when dealing with dysarthric speakers. However, the architecture of metamodels gave rise to difficulties when modelling deletions of sequences of phones, which led us to refine the technique to use weighted finite-state transducers (WFSTs). These were used at the confusion matrix, word, and language levels in a cascade in order to correct errors. The results obtained using this technique were significantly better than those obtained using MLLR, and also better than using metamodels.

The work presented here must be treated as preliminary given the small size of the vocabulary and the restricted syntax of the sentences uttered in the NEMOURS database, and it needs to be validated on a larger dataset with more dysarthric speakers, more utterances per speaker, a larger vocabulary, and a freer syntax. Future work will concentrate on this, and also

(i) applying the techniques described here to normal speech;

(ii) integrating better the confusion matrix transducer with the speech recognizer;

(iii) obtaining robust estimates of confusion matrices from sparse data [47].

References

[1] A. Kain, X. Niu, J. P. Hosom, Q. Miao, and J. van Santen, "Formant re-synthesis of dysarthric speech," in *Proceedings of the 5th ISCA Speech Synthesis Workshop (SSW '04)*, pp. 25–30, Pittsburgh, Pa, USA, June 2004.

[2] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech and Hearing Research*, vol. 12, no. 2, pp. 246–269, 1969.

[3] F. L. Darley, A. E. Aronson, and J. R. Brown, "Clusters of deviant speech dimensions in the dysarthrias," *Journal of Speech and Hearing Research*, vol. 12, no. 3, pp. 462–496, 1969.

[4] R. D. Kent, H. K. Vorperian, J. F. Kent, and J. R. Duffy, "Voice dysfunction in dysarthria: application of the multi-dimensional voice program™," *Journal of Communication Disorders*, vol. 36, no. 4, pp. 281–306, 2003.

[5] R. D. Kent, J. F. Kent, J. Duffy, and G. Weismer, "The dysarthrias: speech-voice profiles, related dysfunctions, and neuropathology," *Journal of Medical Speech-Language Pathology*, vol. 6, no. 4, pp. 165–211, 1998.

[6] W. M. Holleran, S. G. Ziegler, O. Goker-Alpan, et al., "Skin abnormalities as an early predictor of neurologic outcome in Gaucher disease," *Clinical Genetics*, vol. 69, no. 4, pp. 355–357, 2006.

[7] K. Bunton, R. D. Kent, J. F. Kent, and J. R. Duffy, "The effects of flattening fundamental frequency contours on sentence intelligibility in speakers with dysarthria," *Clinical Linguistics & Phonetics*, vol. 15, no. 3, pp. 181–193, 2001.

[8] L. Ramig, "The role of phonation in speech intelligibility: a review and preliminary data from patients with Parkinson's disease," in *Intelligibility in Speech Disorders: Theory, Measurement and Management*, R. D. Kent, Ed., pp. 119–155, John Benjamins, Amsterdam, The Netherlands, 1992.

[9] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 3, pp. 1060–1063, Toulouse, France, May 2006.

[10] H. Strik, E. Sanders, M. Ruiter, and L. Beijer, "Automatic recognition of dutch dysarthric speech: a pilot study," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '02)*, pp. 661–664, Denver, Colo, USA, September 2002.

[11] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[12] P. D. Polur and G. E. Miller, "Effect of high-frequency spectral components in computer recognition of dysarthric speech based on a Mel-cepstral stochastic model," *Journal of Rehabilitation Research and Development*, vol. 42, no. 3, pp. 363–371, 2005.

[13] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *Augmentative and Alternative Communication*, vol. 16, no. 1, pp. 48–60, 2000.

[14] G. K. Poock, W. C. Lee Jr., and S. W. Blackstone, "Dysarthric speech input to expert systems, electronic mail, and daily job activities," in *Proceedings of the American Voice Input/Output Society Conference (AVIOS '87)*, pp. 33–43, Alexandria, Va, USA, October 1987.

[15] N. Thomas-Stonell, A.-L. Kotler, H. A. Leeper, and P. C. Doyle, "Computerized speech recognition: influence of intelligibility and perceptual consistency on recognition accuracy," *Augmentative and Alternative Communication*, vol. 14, no. 1, pp. 51–56, 1998.

[16] P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. Hawley, and M. Parker, "Automatic speech recognition with sparse training data for dysarthric speakers," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech '03)*, pp. 1189–1192, Geneva, Switzerland, September 2003.

[17] A.-L. Kotler and C. Tam, "Effectiveness of using discrete utterance speech recognition software," *Augmentative and Alternative Communication*, vol. 18, no. 3, pp. 137–146, 2002.

[18] L. J. Ferrier, "Clinical study of a dysarthric adult using a touch talker with words strategy," *Augmentative and Alternative Communication*, vol. 7, no. 4, pp. 266–274, 1991.

[19] L. J. Ferrier, H. C. Shane, H. F. Ballard, T. Carpenter, and A. Benoit, "Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition," *Augmentative and Alternative Communication*, vol. 11, no. 3, pp. 165–175, 1995.

[20] A.-L. Kotler and N. Thomas-Stonell, "Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment," *Augmentative and Alternative Communication*, vol. 13, no. 2, pp. 71–80, 1997.

[21] N. J. Manasse, K. Hux, and J. L. Rankin-Erickson, "Speech recognition training for enhancing written language generation by a traumatic brain injury survivor," *Brain Injury*, vol. 14, no. 11, pp. 1015–1034, 2000.

[22] G. Jayaram and K. Abdelhamied, "Experiments in dysarthric speech recognition using artificial neural networks," *Journal of Rehabilitation Research and Development*, vol. 32, no. 2, pp. 162–169, 1995.

[23] C. Goodenough-Trapagnier and M. J. Rosen, "Towards a method for computer interface design using speech recognition," in *Proceedings of the 4th Rehabilitation Engineering and Assistive Technology Society of North America (RESNA '91)*, pp. 328–329, Kansas City, Mo, USA, June 1991.

[24] N. Talbot, "Improving the speech recognition in the ENABL project," *TMH-QPSR*, vol. 41, no. 1, pp. 31–38, 2000.

[25] T. Magnuson and M. Blomberg, "Acoustic analysis of dysarthric speech and some implications for automatic speech recognition," *TMH-QPSR*, vol. 41, no. 1, pp. 19–30, 2000.

[26] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[27] M. S. Hawley, P. Green, P. Enderby, S. Cunningham, and R. K. Moore, "Speech technology for e-inclusion of people with physical disabilities and disordered speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 445–448, Lisbon, Portugal, September 2005.

[28] M. Parker, S. Cunningham, P. Enderby, M. Hawley, and P. Green, "Automatic speech recognition and training for severely dysarthric users of assistive technology: the STARDUST project," *Clinical Linguistics and Phonetics*, vol. 20, no. 2-3, pp. 149–156, 2006.

[29] A. Hatzis, P. Green, J. Carmichael, et al., "An integrated toolkit deploying speech technology for computer based speech training with application to dysarthric speakers," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech '03)*, pp. 2213–2216, Geneva, Switzerland, September 2003.

[30] Clinical Applications of Speech Technology, Speech and Hearing Group, "Voice Input Voice Output Communication Aid (VIVOCA)," Department of Computer Science, University of Sheffield, 2008, http://www.shef.ac.uk/cast/projects/vivoca.

[31] Voicewave Technology Inc., "Speech Enhancer," 2008, http://www.speechenhancer.com/equipment.htm.

[32] J. Rothwell and D. Fuller, "Functional communication for soft or inaudible voices: a new paradigm," in *Proceedings of the 28th Rehabilitation Engineering and Assistive Technology Society of North America (RESNA '05)*, Atlanta, Ga, USA, June 2005.

[33] H. Kim, M. Hasegawa-Johnson, A. Perlman, et al., "Dysarthric speech database for universal access research," in *Proceedings of the International Conference on Spoken Language Processing (Interspeech '08)*, pp. 1741–1744, Brisbane, Australia, September 2008.

[34] Speech Research Lab, A.I. duPont Hospital for Children and the University of Delaware, 2008, http://www.asel.udel.edu/speech/projects.html.

[35] Speech Research Lab, "InvTool Recording Software and ModelTalker Synthesizer," A.I. duPont Hospital for Children and the University of Delaware, 2008, http://www.asel.udel.edu/speech/ModelTalker.html.

[36] X. Menéndez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The Nemours database of dysarthric speech," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, vol. 3, pp. 1962–1965, Philadelphia, Pa, USA, October 1996.

[37] K. Hux, J. Rankin-Erickson, N. Manasse, and E. Lauritzen, "Accuracy of three speech recognition systems: case study of dysarthric speech," *Augmentative and Alternative Communication*, vol. 16, no. 3, pp. 186–196, 2000.

[38] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: a british english speech corpus for large vocabulary continuous speech recognition," in *Proceedings of the 20th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, vol. 1, pp. 81–84, Detroit, Mich, USA, May 1995.

[39] S. Young and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[40] S. J. Cox and S. Dasmahapatra, "High-level approaches to confidence estimation in speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 460–471, 2002.

[41] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '89)*, vol. 1, pp. 532–535, Glasgow, Scotland, May 1989.

[42] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[43] M. Levit, H. Alshawi, A. Gorin, and E. Nöth, "Context-sensitive evaluation and correction of phone recognition output," in *Proceedings of the 8th ISCA European Conference on Speech Communication and Technology (Eurospeech '03)*, pp. 925–928, Geneva, Switzerland, September 2003.

[44] E. Fosler-Lussier, I. Amdal, and H.-K. J. Kuo, "On the road to improved lexical confusability metrics," in *Proceedings of the ISCA Tutorial and Research Workshop on Pronunciation Modelling and Lexicon Adaptation (PMLA '02)*, pp. 53–58, Estes Park, Colo, USA, September 2002.

[45] N. Bodenstab and M. Fanty, "Multi-pass pronunciation adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, vol. 4, pp. 865–868, Honolulu, Hawaii, USA, April 2007.

[46] "Weighted Finite-State Transducer Software Library Lecture," Courant Institute of Mathematical Sciences, New

York University, 2007, http://www.cs.nyu.edu/~mohri/asr07/lecture_2.pdf.

[47] S. J. Cox, "On estimation of a speaker's confusion matrix from sparse data," in *Proceedings of the International Conference on Spoken Language Processing (Interspeech '08)*, pp. 1–4, Brisbane, Australia, September 2008.